CIS6930 Speech Driven Facial Expressions

Rahul Mora
mora.rahul@ufl.edu
4577-9236

Koushik Chanda
koushik.chanda@ufl.edu
2871-9927

Sai Manaswi Kollu
saimanaswikollu@ufl.edu
6961-7116

Swami Charan Rao Beeravelly
s.beeravelly@ufl.edu
2576-1254

ABSTRACT

This research explores the innovative intersection of speech recognition and facial animation, aiming to develop a highly accurate and expressive avatar system capable of generating facial animations directly from audio input. Using advanced machine learning models like LSTM networks, our project bridges the gap between traditional, labor-intensive animation processes and the dynamic capabilities of AI-driven techniques. Through the meticulous analysis of audio files, our system is designed to interpret and visualize speech-driven expressions, converting them into realistic 3D facial motions, including nuanced lip movements and subtle expressions, across a diverse array of speech intensities and patterns.

This approach addresses the significant challenges faced by independent creators in producing animated content with high emotional and expressive fidelity, due to the prohibitive costs and technical expertise required by conventional methods. Our prototype, showcased through an interactive web application, demonstrates the feasibility and efficiency of generating emotionally rich avatars from simple audio inputs, offering a scalable solution that could revolutionize content creation in animation and virtual interactions. Expected outcomes include enhanced lip-sync accuracy, and an improved algorithm for differentiating speech from background noise, thereby ensuring precise facial expression generation in accordance with the spoken content. Through this endeavor, we aim not only to advance the field of speech-driven facial animation but also to democratize the creation of animated content, making it accessible to a broader range of creators and industries.

1 INTRODUCTION

The integration of real-time lip syncing into animation production has revolutionized the creation of believable characters in virtual environments . Traditionally, lip syncing was a labor-intensive post-production task, particularly challenging for multilingual animations due to manual adjustments and time constraints [Taylor et al. 2017].

Creating accurate lip sync animation poses significant difficulties, particularly in mapping lip movements to sound and setting keyframe values manually [Ali et al. 2015]. Traditional methods require meticulous adjustment frame by frame, leading to time-consuming processes, especially for multilingual animations. Real-time approaches offer a solution to these challenges by automating the process of lip sync animation, reducing the need for manual adjustments, and ensuring realism by mapping key phoneme sounds to corresponding lip shapes. This approach streamlines the animation production process, shortens production durations, and ensures accurate lip sync outcomes, particularly for animations intended for broadcast in multiple languages. Achieving realistic character animation requires accurately matching the lip movements to the spoken sounds, which is crucial for creating a convincing visual experience [Taylor et al. 2017].

Integration of real-time lip sync animation into existing animation production pipelines offers numerous benefits, such as ease of implementation in synthesis phases and applicability to various animator tasks, including multilingual animation reproduction, pre-animation production, and avatar speech in gaming pipelines. By leveraging audio signals in real-time, this approach enables efficient and accurate lip sync animation, enhancing the overall quality and realism of animated content across diverse applications and industries.

Believable characters or agents in interactive systems serve various purposes, from personal assistants to educational tools, leveraging familiar visual interfaces like human faces to enhance user engagement. Faces are effective in communication due to their ability to convey emotions [Ekman 1989], making them valuable in applications such as information kiosks and educational simulations. Automation of facial animation has been a focus in animation and gaming industries [Thalmann and Thalmann 2002], with efforts to create realistic expressions through advanced algorithms and models. However, integrating emotional responses into dynamic environments remains a challenge, requiring the connection of facial expression generation with computational models of emotions, often drawing from psychological principles, and learning algorithms.

With inputs from psychologists, many contributions were made in the Intelligent Agents' field to create models of emotions [Bates 1994]. While some models focus on mapping events to emotional responses based on their impact on the agent's objectives [Ortony et al. 1988], others prioritize internal states like fatigue, pain, and thirst. Learning has been identified as a crucial element in creating believable behaviors in dynamic environments, enabling agents to recognize action patterns, form associations with positive or negative events, adapt to repeated actions, and overall adjust to their surroundings. In our research, a novel model of emotions and lip movements for intelligent agents has been developed, utilizing multiple learning algorithms to facilitate the generation of realistic behavior.

This paper presents an integration of facial expression models with lip movements models to create an avatar that generates expressiveness in real-time interactions. Our model updates the avatar's facial expressions based on a given audio, generating behaviors consistent with its experiences, and rendering appropriate facial expressions accordingly. This approach mirrors the process used by computer animators in character development, coupling emotional response algorithms with facial expression models to create powerful character design methodologies with potential applications in interactive environments, such as video games.

2 RELATED WORK

The pursuit of realistic and emotionally expressive avatars through the integration of speech recognition and animation technologies has seen significant advancements, guided by a rich body of research spanning across machine learning, computer vision, and interactive digital media. This literature survey delves into key studies that have paved the way for our project, highlighting the evolution of techniques from manual animation to AI-driven processes.

In 'CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior,' Xing et al. (2023) tackle the considerable challenge of enhancing the realism and vividness in speech-driven 3D facial animation, confronting the issues arising from the complex nature of audio-visual data mapping and limited data availability. They propose an inventive solution by reimagining the animation process as a code query within a defined proxy space, leveraging a codebook imbued with genuine facial motion priors obtained through self-reconstruction of real facial movements. This unique strategy allows for the use of a temporal autoregressive model to create facial animations from speech inputs, achieving remarkable lip sync accuracy and believable expressions. The effectiveness of their methodology is underscored by noticeable improvements in the realism and dynamism of animations compared to existing techniques, both in qualitative and quantitative terms. Additionally, a user study highlights the superior perceptual quality of animations generated using their approach, affirming its potential to significantly advance the field of facial animation [Xing et al. 2023].

In the paper, "Emotion-Aware Transformer Encoder for Empathetic Dialogue Generation," Goel, Susan, Vashisht, and Dhanda (2021) present a cutting-edge method that seamlessly integrates emotional sensitivity into the conversational model's learning process. This approach features an innovative emotion detector that evaluates the emotional state from the user's speech, enabling the system to tailor its responses with enhanced empathy and human-like qualities. Additionally, their pioneering transformer encoder merges word and emotion embeddings, infusing the dialogue with a rich emotional layer. Utilizing the advanced Transformer-XL architecture, their approach outperforms existing models on the Facebook AI empathetic dialogue dataset, achieving higher BLEU-4 scores. This breakthrough highlights the significant role emotionally intelligent agents can play in elevating the quality of human-machine interactions, paving the way for more emotionally engaging technologies soon [Goel et al. 2021].

Thambiraja et al. (2022) introduce "Imitator," a novel method for creating personalized speech-driven 3D facial animations that capture the unique speaking styles and facial idiosyncrasies of individuals. Unlike previous techniques that often neglect the actor's specific facial characteristics, leading to unrealistic lip movements, Imitator leverages a style-agnostic transformer trained on a vast dataset of facial expressions. This approach allows for the generation of facial expressions that are not only in sync with the audio input but also reflect the distinct speaking mannerisms of the target actor. A key innovation of their work is the development of a loss function that focuses on the accurate reproduction of bilabial consonants, enhancing the realism of lip movements. Through comprehensive testing and a user study, the team demonstrates that Imitator effectively produces temporally coherent facial animations that maintain the authenticity of the actor's own speaking style [Thambiraja et al. 2022].

3 METHOD

Our project develops an interactive web-based platform that dynamically generates expressive avatars reacting to audio inputs. The system architecture is divided into two primary components: the frontend user interface developed with modern JavaScript frameworks and the backend processing pipeline powered by Python.

3.1 Overall Architecture

At the core of our system lies an intricately designed architecture that seamlessly integrates cutting-edge web technologies with advanced machine learning algorithms.

The frontend, crafted using React.js, offers a user-friendly interface for audio file upload and real-time avatar interaction, ensuring a smooth and engaging user experience. This frontend communicates with the backend through RESTful APIs, facilitating the efficient transmission of audio data for processing.

On the backend, a Flask-based server handles these requests, employing Python's powerful libraries to preprocess the audio and prepare it for sentiment analysis. This symbiotic relationship between the frontend and backend not only optimizes performance but also enables the real-time processing essential for live avatar animation.

3.2 Machine Learning Models for Sentiment Analysis

Our project leverages deep learning techniques to analyze audio inputs and predict the sentiment expressed, enabling the generation of corresponding facial expressions in digital avatars. We utilize the TensorFlow framework to develop and train our sentiment analysis model, with specifics as follows:

3.2.1 ***Model Architecture.*** The core of our sentiment analysis is based on a Convolutional Neural Network (CNN) combined with Recurrent Neural Network (RNN) layers, specifically utilizing Long Short-Term Memory (LSTM) units [Hung and Tien 2021]. This architecture is adept at processing sequential data, making it ideal for audio analysis where temporal relationships within the signal are crucial for understanding emotional content [Satt et al. 2017].

3.2.2 ***Training Data.*** The model is trained on a curated dataset comprising audio clips labeled with emotional sentiments, including happiness, sadness, anger, and neutrality. Each audio clip is preprocessed into Mel-frequency cepstral coefficients (MFCCs) using the librosa library, providing a time-frequency representation of the sound that serves as input to the model [Schuller et al. 2013].

3.2.3 ***Parameters and Configuration.***

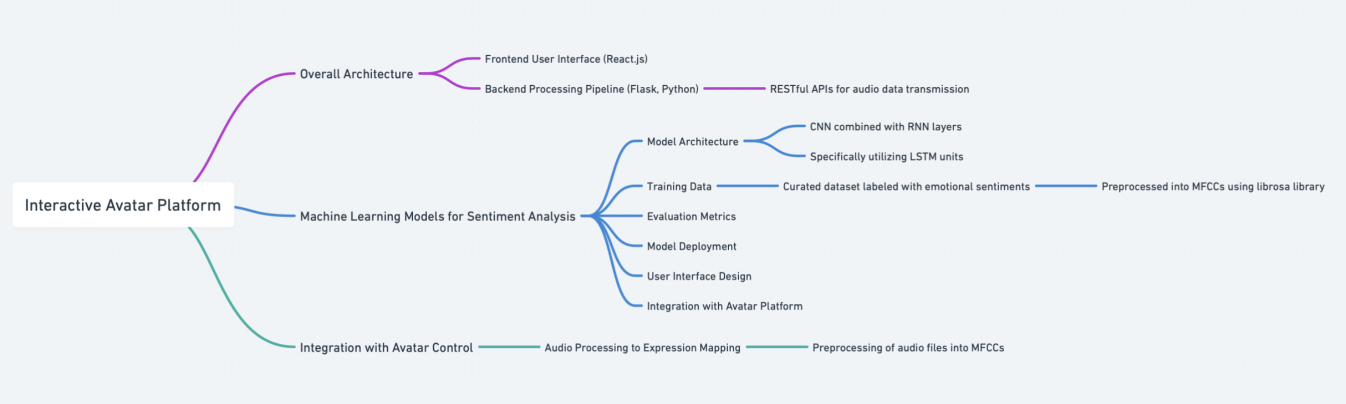

Interactive Avatar Platform
Overall Architecture
Frontend User Interface (React.js)
Backend Processing Pipeline (Flask, Python)
RESTful APIs for audio data transmission
Machine Learning Models for Sentiment Analysis
Model Architecture
CNN combined with RNN layers
Specifically utilizing LSTM units
Training Data
Curated dataset labeled with emotional sentiments
Preprocessed into MFCCs using librosa library
Evaluation Metrics
Model Deployment
User Interface Design
Integration with Avatar Platform
Integration with Avatar Control
Audio Processing to Expression Mapping
Preprocessing of audio files into MFCCs

Figure 1: System architecture with all the stages of workflow

MFCCs: 13 coefficients, capturing the primary audio characteristics.

CNN Layers: Two sets of convolutional layers, the first comprising 32 filters and the second 64 filters, are succeeded by max-pooling layers to decrease dimensionality.

LSTM Units: 128 units, capturing long-term dependencies in audio sequences.

Dense Layer: A densely connected layer using a softmax activation function is utilized to categorize the audio into emotional classifications.

Output Layer: A softmax layer that classifies the emotion into categories such as happy, sad, or neutral based on the features recognized by the network.

Training Process: We employed an 80-20 split to allocate data for both training and validation purposes. We employed categorical cross-entropy as the loss function and utilized the Adam optimizer to adjust the weights during the training process.

3.3 Integration with Avatar Control

The integration of the sentiment analysis model with avatar control is achieved through a mapping between the predicted emotional categories and predefined facial expressions. The integration process begins with the audio input from the user, which undergoes preprocessing to extract meaningful features (MFCCs) that are then fed into our sentiment analysis model. Based on the emotion predicted, a corresponding set of facial expressions is generated on the avatar.

To integrate lip syncing with the dynamic expression generation in the avatar, we expanded our system's capabilities to include real-time lip movement synchronized with the audio playback. This enhancement ensures that the avatar not only displays emotional expressions but also accurately mimics speech movements, providing a more lifelike and engaging interaction.

3.3.1 Audio Processing to Expression Mapping.

Preprocessing: Audio files uploaded by the user undergo a transformation into Mel-frequency cepstral coefficients (MFCCs), a representation that effectively captures the essential characteristics of the audio for emotion detection. These characteristics include timbral texture, pitch, tone, and dynamics, which are crucial in differentiating emotional states in spoken language. The MFCCs serve as a compact representation of these audio features, providing our sentiment analysis model with a robust input that highlights the nuances in speech that correspond to different emotions, thereby enhancing the accuracy of emotion prediction.

Model Prediction: The preprocessed audio data is passed through the CNN-Bi-LSTM model, which predicts the primary emotional sentiment expressed in the audio clip.

Emotion to Expression Translation: Upon predicting an emotion from the audio, the system translates this emotional sentiment into facial expressions through a predefined mapping between emotions and specific facial animation parameters, known as blend shapes. These blend shapes are meticulously designed to accurately represent various facial expressions associated with emotions such as happiness, sadness, and anger. The translation process involves activating the appropriate blend shapes to varying degrees, allowing for nuanced expressions that reflect the intensity and complexity of the detected emotion. This dynamic adjustment of blend shapes ensures that the avatar's expressions change in real-time to mirror the emotional content of the spoken words, creating a visually expressive and engaging interaction.

3.3.2 ***Lip Syncing Implementation.*** The incorporation of lip syncing into our avatar control system involved analyzing the phonetic content of the audio input to determine the appropriate mouth shapes, or visemes, corresponding to spoken sounds. This process is executed as follows:

Phonetic Analysis: Utilizing advanced speech recognition technologies, the system breaks down the audio input into phonetic components. Each phoneme, the smallest unit of sound in speech, is mapped to a specific viseme that represents how the mouth should look when making that sound. A database of visemes associated with various phonemes guides the animation of the avatar's mouth.

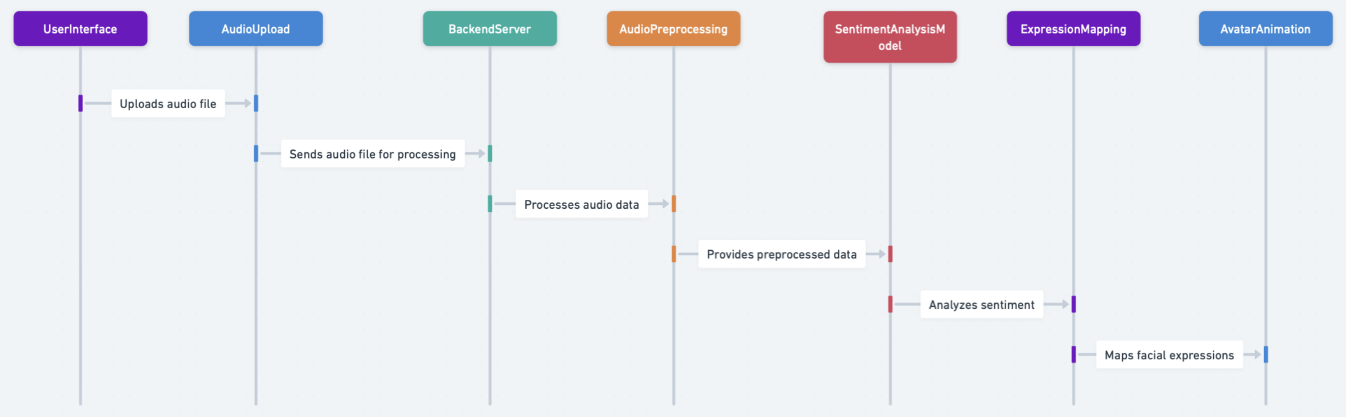


Figure 2: Block Diagram of the workflow

These visemes are carefully crafted blend shapes within the avatar's model, designed to represent the full range of human mouth movements during speech.

Real-Time Synchronization: As the audio plays, the system dynamically updates the avatar's mouth position in real time, synchronizing lip movements with the audio's phonetic sequence. This synchronization is achieved through precise timing controls that match viseme transitions to the audio waveform, ensuring that the avatar's lips move in concert with the spoken words.

Integration with Emotional Expressions: The lip syncing mechanism operates in tandem with the emotional expression generation, allowing the avatar to simultaneously convey spoken content and emotional nuances. Special attention is given to maintaining natural transitions between visemes and expressions, preserving the continuity and realism of the avatar's facial movements.

3.3.3 ***Dynamic Expression Generation****.* Upon audio input, the backend processes the file and feeds the extracted features into the sentiment analysis model. The model's output is then transmitted to the frontend, where JavaScript functions interpret the emotional category and dynamically adjust the avatar's facial blend shapes in real time.

This combined approach of dynamic expression generation and lip syncing significantly enhances the avatar's expressiveness and realism. By accurately mimicking both the emotional and verbal aspects of human communication, the avatar provides users with a richly interactive experience, bridging the gap between digital and real-world interactions.

This approach combines cutting-edge machine learning with sophisticated avatar animation techniques, resulting in expressive digital agents capable of reflecting a wide range of human emotions. It represents a significant advancement in the field of interactive technology, offering new possibilities for enhancing virtual communication and storytelling.

4 RESULTS

Our investigation into utilizing LSTM configurations for training avatars in expression matching on the TIMIT dataset has yielded substantial results. The Bidirectional LSTM configuration emerged as the most effective, demonstrating superior capability in understanding and replicating complex expressions. It achieved a notable training accuracy of 64.5% and a testing accuracy of 62.3%. Specifically, this model excelled in phoneme to expression matching, with an accuracy of 66.9%, and viseme to expression matching, where it reached an accuracy of 77.1% (shown in Figure : 3). These outcomes underscore the model's adeptness at capturing the nuanced temporal dynamics and dependencies inherent in speech, which are crucial for accurate expression replication in avatars.

In contrast, the standard LSTM model [Li et al. 2019], while still delivering respectable performance, reported slightly lower accuracies: 54.9% in training and 55.4% in testing. The absence of detailed accuracies for phoneme and viseme expression matching in the full test set for this model indicates the need for further exploration to fully understand its performance compared to the bidirectional approach.

The final output is depicted in Figure 4

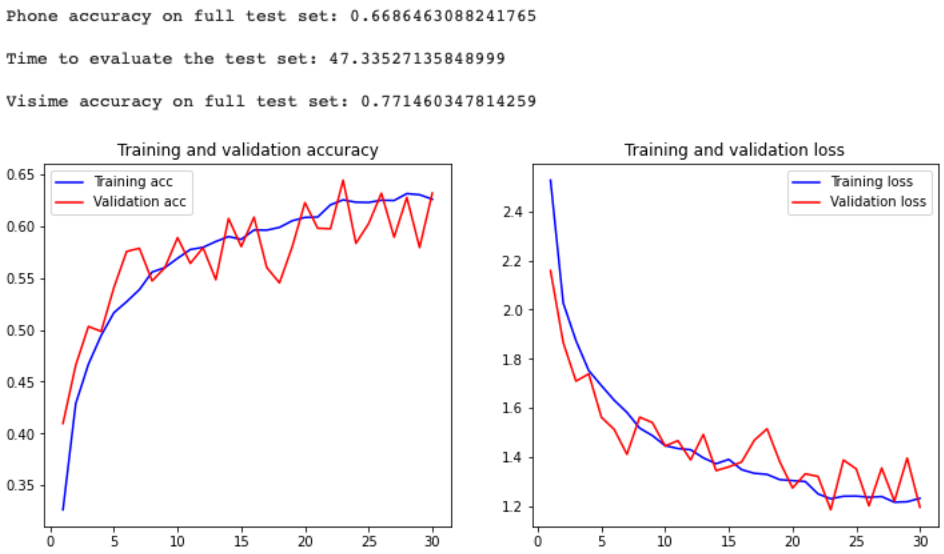
The detailed training logs provided valuable insights into the progression of model accuracies over epochs, revealing the impact of hyperparameters such as dropout rates and LSTM units on learning trajectories. This information is vital for unraveling model behaviors and steering future optimizations in avatar expression training technologies.

5 LIMITATIONS

Despite the promising advancements, several limitations were encountered throughout the study. First, the complexity of human expressions and their nuances presents a significant challenge, indicating that even the most advanced models have room for improvement to achieve truly life-like avatar expression matching.

Secondly, the exclusive use of the TIMIT dataset, while beneficial for consistency and benchmarking, may not fully represent the diversity of expressions across different languages and cultures. This limitation suggests the potential benefits of incorporating a broader range of datasets for more generalized and robust avatar training.

Thirdly, the computational demands, especially of the Bidirectional LSTM models, highlight a significant consideration for real-world applications. The resource-intensive nature of these models



```
Phone accuracy on full test set: 0.6686463088241765

Time to evaluate the test set: 47.33527135848999

Visime accuracy on full test set: 0.771460347814259
```

Figure 3: Training Accuracy and loss for Bi-directional LSTM

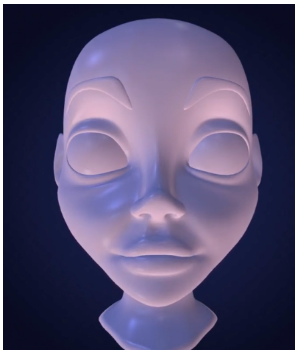


Figure 4: Final Output

could limit their deployment in environments with constrained computational capacities.

Finally, this project focused predominantly on LSTM-based models, not considering the potential applicability of recent advancements in neural network architectures, such as Transformers, which might offer further improvements in expression matching tasks.

6 FUTURE WORK

Future work for the model can focus on many key aspects to enhance its capabilities and effectiveness. Firstly, extending the model's proficiency to accurately generate lip movements and expressions for different languages would significantly broaden its applicability and accessibility across diverse linguistic contexts. This involves training the model on a more extensive dataset encompassing various languages and dialects, as well as refining the mapping of phoneme sounds to corresponding lip shapes for each language [El-Bialy et al. 2022]. Additionally, incorporating cultural nuances and speech patterns specific to different languages can further improve the authenticity and naturalness of the generated animations.

We can further refine the model's algorithms to improve the realism and naturalness of the generated lip movements and facial expressions. This could involve incorporating advanced techniques such as physics-based simulations or neural rendering to achieve more lifelike animations.

We tend to address ethical considerations related to the deployment of expressive agents, such as privacy, bias, and inclusivity. We can implement safeguards to ensure that the model respects user privacy, mitigates bias in its responses.

Additionally, exploring applications beyond lip-syncing, such as full-body animation or gesture recognition, could extend the utility of our model to a wider range of interactive scenarios. This could involve developing multi-modal models that integrate audio, visual, and gestural inputs to create more immersive and interactive virtual characters.

Finally, integrating a feedback mechanism into the model to iteratively improve its performance based on user input and evaluations is crucial. Gathering feedback from users, animators, and domain experts can provide valuable insights into areas for enhancement and refinement. This feedback loop can be utilized to fine-tune the model's algorithms, optimize its parameters, and address any discrepancies observed in the generated lip movements and expressions.

REFERENCES

Itimad Ali, Ghazali Sulong, and Hoshang Kolivand. 2015. Realistic Lip Syncing for Virtual Character Using Common Viseme Set. *Computer and Information Science* 8 (08 2015). https://doi.org/10.5539/cis.v8n3p71

Joseph Bates. 1994. The role of emotion in believable agents. *Commun. ACM* 37, 7 (jul 1994), 122–125. https://doi.org/10.1145/176789.176803

Paul Ekman. 1989. The argument and evidence about universals in facial expressions of emotion. https://api.semanticscholar.org/CorpusID:202243798

Randa El-Bialy, Daqing Chen, Souheil Fenghour, Walid Hussein, Perry Xiao, Omar H. Karam, and Bo Li. 2022. Developing phoneme-based lip-reading sentences system for silent speech recognition. *CAAI Transactions on Intelligence Technology* 8, 1 (Aug. 2022), 129–138. https://doi.org/10.1049/cit2.12131

Raman Goel, Seba Susan, Sachin Vashisht, and Armaan Dhanda. 2021. Emotion-Aware Transformer Encoder for Empathetic Dialogue Generation. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. https://doi.org/10.1109/aciiw52867.2021.9666315

Bui Thanh Hung and Le Minh Tien. 2021. *Facial Expression Recognition with CNN-LSTM.* Springer Singapore, 549–560. https://doi.org/10.1007/978-981-15-7527-3_52

Tzuu-Hseng S. Li, Ping-Huan Kuo, Ting-Nan Tsai, and Po-Chien Luan. 2019. CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot. *IEEE Access* 7 (2019), 93998–94011. https://doi.org/10.1109/access.2019.2928364

Andrew Ortony, Gerald Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotion.* Vol. 18. https://doi.org/10.2307/2074241

Avner Satt, Shai Rozenberg, and Ronen Hoory. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech.* 1089–1093.

Bj"orn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Interspeech.* 148–152.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Trans. Graph.* 36, 4, Article 93 (jul 2017), 11 pages. https://doi.org/10.1145/3072959.3073699

Nadia Thalmann and Daniel Thalmann. 2002. Computer Animation. *Comput. Surveys* 28 (05 2002). https://doi.org/10.1145/234313.234381

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. 2022. Imitator: Personalized Speech-driven 3D Facial Animation. arXiv:2301.00023 [cs.CV]

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. arXiv:2301.02379 [cs.CV]